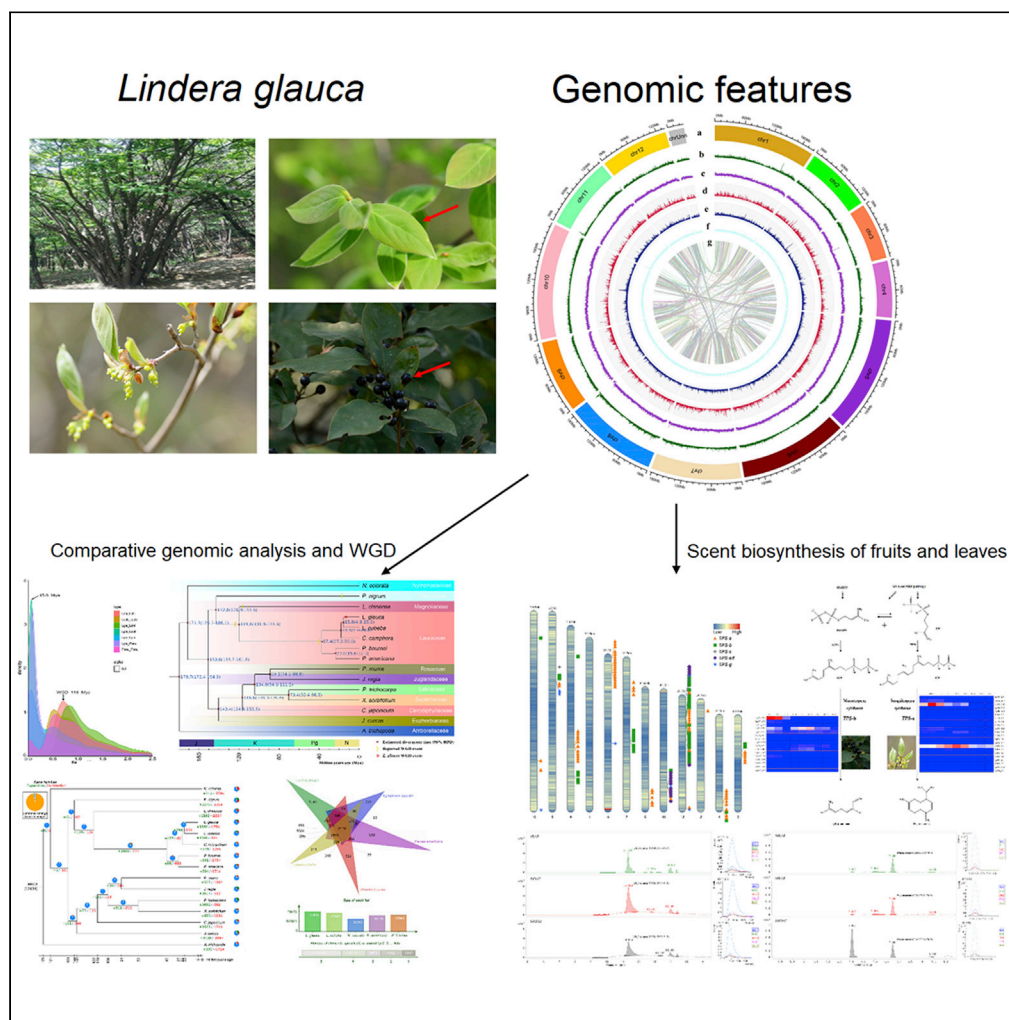


Article

Genome of *Lindera glauca* provides insights into the evolution of biosynthesis genes for aromatic compounds

Biao Xiong, Limei Zhang, Lun Xie, ..., Shuai Liao, Shubin Dong, Zhixiang Zhang

zxzhang@bjfu.edu.cn (S.D.)
dongshubin@bjfu.edu.cn (Z.Z.)

Highlights

We provide the first chromosome-level genome for *Lindera glauca*

We explore the phylogenetic position and identify three WGD events of *L. glauca*

We identify genes involved in the main aromatic compounds

Analysis of metabolites in fruits and leaves by GC-MS is reported

Xiong et al., iScience 25, 104761
August 19, 2022 © 2022 The Author(s).
<https://doi.org/10.1016/j.isci.2022.104761>

Article

Genome of *Lindera glauca* provides insights into the evolution of biosynthesis genes for aromatic compounds

Biao Xiong,^{1,2,6} Limei Zhang,² Lun Xie,¹ Linzhu Li,³ Xiangxiang He,⁴ Yan Niu,⁴ Tianyuan Zhang,⁴ Shuai Liao,⁵ Shubin Dong,^{2,*} and Zhixiang Zhang^{2,*}

SUMMARY

***Lindera glauca* is a crucial source of diverse industrial oil and medicines. The spicy aroma of tender leaves is caused by the presence of abundant aromatic compounds. Here, we present its chromosome-level genome assembly comprising 12 pseudochromosomes (2,092.2 Mb; scaffold N50: 186.5 Mb), which was predicted to have 65,145 protein-coding genes. Comparative genomic analyses indicated two whole-genome duplication (WGD) events in the Lauraceae family, contributing to the production of numerous terpene synthase (TPS) genes. We identified 138 TPS genes in *L. glauca*. Comparative transcriptomic analyses revealed high expression of genes *Lg03G2346* and *Lg08G140* in TPS-a and *Lg07G2961* and *Lg12G971* in TPS-b subfamilies, which regulated the biosynthesis of the monoterpenoid β -ocimene and sesquiterpenoid D-germacrene in *L. glauca*. The results suggested a molecular basis for species-specific terpenoid biosynthesis and provided a foundation for molecular breeding to produce desired characteristics and a valuable reference genome.**

INTRODUCTION

L. glauca, a deciduous shrub or small tree of the Lauraceae family, is widely distributed in low-altitude montane forests of central and southern China as well as in Japan, Korea, and Vietnam and has been used medicinally for centuries (Tsuji and Werff, 2008). It contains compounds such as terpenoids, flavonoids, and alkaloids that have anti-inflammatory (Ruan et al., 2020), antioxidative (Kim et al., 2019), antiviral (Park et al., 2018), and antitumor activities (Kim et al., 2014; Suh et al., 2015). The fruits of *L. glauca* are a rich source of fatty acids and aromatic oils, which are used as raw materials to produce lubricants and biodiesel (Wang et al., 1994; Qi et al., 2016; Xiong et al., 2018). Its leaves are used to prepare a traditional folk beverage or in Chinese medicine because of their high terpenoid content (Sun et al., 2011; Zhu et al., 2016). *L. glauca* is especially popular in parts of southwest China because of the special scents produced by its leaves.

The aromatic compounds produced by many species of the Lauraceae family have various components, primarily terpenoids or their derivatives. Terpenoids constitute the largest class of structurally diverse metabolites, with more than 80,000 members found in numerous living organisms including plants (Christianson, 2017). They play crucial roles in the physiological functions of plants, including growth and development, environmental adaptation, pollinator attraction, and resistance to pests, diseases, and heat (Nagegowda, 2010). Terpenoids are widely used in cosmetics (e.g., limonene for producing essential oil) and as medicines, such as the anticancer drugs paclitaxel, oleuropein, and ginsenoside (Bulotta et al., 2011; Gutensohn et al., 2013); artemisinin for treating malaria; and tanshinone for treating coronary heart disease. Terpenoids are classified into monoterpenes (C₁₀), sesquiterpenes (C₁₅), diterpenes (C₂₀), triterpenes (C₃₀), tetraterpene (C₄₀), and polyterpenes, according to the number of five-carbon units present in their chemical structures (Nagegowda and Gupta, 2020). Terpenoids including monoterpenes, sesquiterpenes, and diterpenes, are abundant in the leaves, flowers, fruits, and seeds of *L. glauca*. In particular, the fruits and leaves have the highest content; their extracts contain 52.96 and 15.17% monoterpenes and 42.16 and 76.14% sesquiterpenes, respectively (Sun et al., 2011). The most abundant terpenoid in *L. glauca* fruits is (Z)-3,7-dimethyl-1,3,6-octatriene (31.90%) and in *L. glauca* leaves is D-germacrene (45.56%) (Sun et al., 2011).

¹College of Tea Science, Guizhou University, Guiyang, China

²Beijing Forestry University, Beijing, China

³Guizhou Academy of Testing and Analysis, Guiyang, China

⁴Wuhan Benagen Technology Co., Ltd.

⁵East China Normal University, Shanghai, China

⁶Lead contact

*Correspondence: zxzhang@bjfu.edu.cn (S.D.), dongshubin@bjfu.edu.cn (Z.Z.)

<https://doi.org/10.1016/j.isci.2022.104761>



In plants, dimethylallyl pyrophosphate and isopentenyl diphosphate are the common precursors of terpenoids, and terpenoid synthesis occurs through the following pathways (Nagegowda and Gupta, 2020): the mevalonic acid (MVA) pathway, which is found in the cytoplasm of eukaryotes, and the methylerythritol phosphate (MEP) pathway, which occurs in the plastids of prokaryotes and plants (Cordoba et al., 2009). Isopentenyl pyrophosphate (IPP) is the central precursor in both the MVA and MEP pathways (Enfissi et al., 2005). The initial precursors in terpenoid biosynthesis are acetyl-CoA, pyruvate, and D-glyceraldehyde-3-phosphate, which are produced in the pentose phosphate pathways (PPP) and Embden-Meyerhof-Parnas (EMP) pathways (Tholl, 2006) and are subsequently involved in both the MVA and MEP pathways. In the MVA pathway, the following six enzymes are directly involved in the conversion of acetyl-CoA thiolase (ACOT) to IPP in the given order; the sequential action of the enzymes is as follows: ACOT, 3-hydroxy-3-methylglutaryl-coenzyme A synthase (HMGS), HMG-CoA reductase (HMGR), mevalonate kinase, phosphomevalonate kinase, and mevalonate diphosphate decarboxylase (Liao et al., 2014; Nagegowda and Gupta, 2020). In the MEP pathway, seven enzymes catalyze pyruvate and D-glyceraldehyde-3-phosphate conversion in seven steps to produce IPP and dimethylallyl diphosphate (DMAPP): 1-deoxy-D-xylulose-5-phosphate synthase, reductoisomerase, 2-C-methyl-D-erythritol 4-phosphate cytidyltransferase, 4-(cytidine 5-diphospho)-2-C-methyl-D-erythritol kinase, 2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase, (E)-4-hydroxy-3-methylbut-2-enyl diphosphate synthase, and (E)-4-hydroxy-3-methylbut-2-enyl diphosphate reductase (Rohdich et al., 2003; Nagegowda and Gupta, 2020).

The two pathways (MVA and MEP) are mainly divided into three stages (Nagegowda and Gupta, 2020). In the first stage, IPP and its allylic DMAPP are synthesized. The second stage involves the synthesis of three direct precursors: geranyl diphosphate (GPP), farnesyl diphosphate (FPP), and geranylgeranyl diphosphate (GGPP). The third stage involves the formation of terpenoids and their structural modification. Monoterpenes, sesquiterpenes, and diterpenes are formed by the condensation of C5 precursors by activating terpene synthases (TPSs) (Tholl, 2006; Chen et al., 2011). TPSs catalyze the biosynthesis of monoterpenes and sesquiterpenes and induce TPS expression (Zulak et al., 2009). Chen et al. (2020a) reported TPS gene families in Lauraceae and the genome of *Litsea cubeba*; however, how these genes and aroma production are interlinked remain unclear. Moreover, knowledge of the evolution and diversification of TPS genes in some crucial economic plants, such as *L. glauca*, remains limited because of limited data on the Lauraceae genome. Considering the chemical polymorphism evident in *L. glauca* terpenoids, identifying genes related to TPSs and searching for potential differences in the biosynthesis pathway may yield valuable findings.

Advancements in genomic analysis have markedly enhanced the identification of genes related to biosynthesis and sequence evolution (Hu et al., 2019; Pu et al., 2020). To elucidate the evolution of biosynthesis genes for aroma compounds, we present a high-quality *de novo* chromosome-level genome of male *L. glauca* trees ($2n = 2x = 24$) obtained using Oxford Nanopore Technologies sequencing (ONT, Oxford, UK), next-generation sequencing (Illumina, USA), and Hi-C technologies (Chromosome conformation capture, 3C). Genomic data, transcriptome, and gas chromatography-mass spectrometry (GC-MS) were used to analyze the contents and composition of terpenoids in leaves and fruits and the genes involved in the MEP/MVA pathway and terpenoid synthesis in the two chemical types. Our findings can contribute to biological and agronomic research in *Lindera* species and other Lauraceae.

RESULTS

Genome assembly and main features

On the basis of *K*-mer genome survey analysis, *L. glauca* was estimated to have a genome size of 2,215.47 Mb. The length of 41-mer analysis indicates that the genome had high heterozygosity (1.42%) and a repetitive sequence content of 76.75% (Figure S1 and Table S1).

To mitigate the effects of high heterozygosity and repetitive sequence content on the assembly of a chromosome-scale reference genome, a comprehensive *de novo* assembly strategy combining approximately 252.4 Gb (113.7× coverage) of raw ONT reads (read N50 = 20.2 kb; contig N50: 710.7 kb), 218.7 Gb of raw Illumina paired-end reads (98.5× coverage), and 129 Gb of Hi-C clean data (58.2× coverage) was performed. Because of the unstable quality of ONT raw reads, the sequenced reads were first corrected twice by using RACON, and 215.3 Gb of corrected reads were obtained (N50:22.3 kb) These reads were then polished twice based on Illumina short reads by using Pilon. The DNA mapping rate was 99.8%

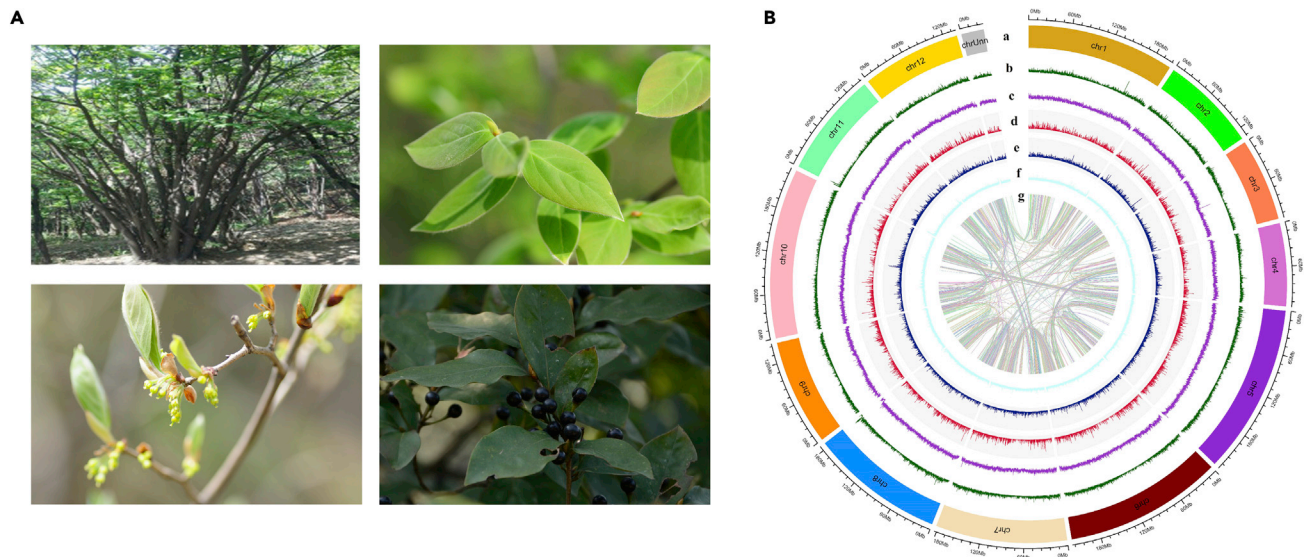


Figure 1. Morphology and genome features of *L. glauca*

(A) *L. glauca*: individual, tender leaves, female inflorescence, and fruits (from top left to bottom right).

(B) Landscape of genome assembly and annotation (chr1–chr12, chr Unn: contigs were unmapped to chromosomes): (a) Chromosome-scale pseudomolecules (chr1–chr12, chr Unn). (b–f) Distributions of gene density, repeat density, copia and gypsy LTR density, and GC density, respectively. (g) Syntenic blocks.

and the complete genome assembly was 94.2% based on the Benchmarking Universal Single-Copy Orthologs (BUSCO) assessment, indicating a high degree of genomic assembly integrity. (Figures 1 and S14, Tables S2, S3, and S4). Using Hi-C sequence data, we anchored scaffolds into 12 pseudomolecules (Figure S14 and Table S5), which improved scaffold N50 to 186.5 Mbp, with the largest scaffold being 236.3 Mb (Table 1). The final chromosome-level genome of *L. glauca* was 2,092.2 Mb (scaffold N50: 186.5 Mb), accounting for 94.4% of the estimated genome size.

With the largest genome size among the published Lauraceae genomes (Chaw et al., 2019; Chen et al., 2020b) (Table S7), the *L. glauca* genome was composed of 62.78% (1,313.4 Mb) transposable elements and 35.53% (743.5 Mb) long terminal repeat (LTR) retrotransposons (Figure 1 and Table S6). Thus, LTR gypsy and copia elements were found to contribute considerably much to the large size of the *L. glauca* genome (Table 1 and Table S6).

A total of 70,484 *L. glauca* genes were predicted, and BUSCO assessments indicated that the completeness of genome annotation was 92.3% (Table S8). After nonfunctional annotations were removed, a high-confidence set of 65,145 protein-coding genes were predicted, with an average coding sequence length of 832 nt. We annotated 37,097 (56.95%), 42,253 (64.86%), 36,659 (56.27%), 20,256 (31.09%), 12,736 (19.55%), 51,311 (78.76%), and 54,394 (83.50%) of the genes of *L. glauca* by using Uniprot, Pfam, Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG), Pathway, Interproscan, and NR, respectively (Table S9). Finally, 59,646 (91.56%) of 65,145 genes were annotated. In particular, 3,550 noncoding RNAs (ncRNAs) were identified, including 496 rRNAs and 858 tRNAs.

Comparative genomic and phylogenomic analyses

An analysis of genomic synteny using MCScanX revealed 13,724 colinear blocks across the whole *L. glauca* genome, which included 18,250 (25.89%) of the genes. Among these syntenic blocks, 6,412 (46.7%) of the paralogous gene pairs were located interchromosomally, and 7,312 (53.3%) were within chromosomes. In addition, 1,280 colinear blocks and 11,315 colinear gene pairs were detected between *Vitis vinifera* and *L. glauca*, and 1,339 and 13,329, respectively, between *V. vinifera* and *L. cubeba* (Figures 2A, 2C, S8, and S9). Two-to-three and one-to-two syntenic relationships were observed when comparing *L. glauca* with *V. vinifera* (Figure 2A), and the same relationship was observed between *L. cubeba* with *V. vinifera* (Figure S8), suggesting similarity in the genome features of *L. glauca* and *L. cubeba*.

Table 1. Major indicators of the *Lindera glauca* genome

Assembly feature	Statistic
Estimated genome size (by 41-mer analysis) (Mb)	2,215.47
Contig N50 (bp)	710,747
Contig N90 (bp)	183,155
Scaffold N50 (Mb)	186.54
Scaffold N90 (Mb)	115.15
Longest scaffold (Mb)	236.26
Assembled genome size (Mb)	2,092.24
GC content	39.62%
Assembly % of genome	94.44%
Predicted gene models	70,484
Average coding sequence length (bp)	4,347.60
Average exons per gene	3.78

Gene order analysis revealed colinear regions with up to five (but mostly two or three) paralogous segments in the *L. glauca* genome (Table S16). The calculated synonymous substitution rate (K_S) of these gene pairs peaked at 0.14–0.19 was based on the slow substitution rate of basal angiosperms. The distribution of synonymous substitutions per K_S for all paralogous genes indicated three clear peaks (Figures 3A and S10): $K_S \approx 0.08$, $K_S \approx 0.52$, and $K_S \approx 0.88$. The first K_S peak (0.08) corresponds to a recent polyploidy event, then the later two K_S peaks also correspond to a polyploidy event, suggesting that *L. glauca* experienced three polyploidy events. The analysis collinear point diagram (Figure 2B) also confirmed that these three polyploidy events of *L. glauca* were WGD events. The K_S peak values for the two earlier ancient whole-genome duplication (WGD) events were nearly identical in *L. glauca*, *L. cubeba* (Chen et al., 2020c), and *Phoebe bournei* (Chen et al., 2020a), suggesting that the two WGD events were common to these taxa (Lauraceae). The two ancient K_S peaks were greater than these differentiation K_S peaks of *L. cubeba*, *P. bournei*, and *L. glauca* (Figure 3A), indicating two ancient WGD events before the common ancestor of Lauraceae diverged. The differentiation peak ($K_S \approx 0.83$) between *L. glauca* and *Liriodendron chinense* was greater than the second peak ($K_S \approx 0.52$) and smaller than the third peak ($K_S \approx 0.88$) for the ancient WGD event in *L. glauca*, suggesting that the most ancient WGD event occurred in the *L. glauca* genome before the differentiation of *L. chinense* (Magnoliales) and Lauraceae and that the second event occurred after their differentiation. These results are consistent with those of previous studies on *L. cubeba* (Chen et al., 2020b), *L. chinense* (Chen et al., 2019), and *Persea americana* (Reñón-Anaya et al., 2019). The recent WGD peaks of Lauraceae species (Figure S10) occurred shortly after the divergence of Lauraceae, implying that these WGD events might have facilitated the rapid radiation of this family (Figures 3A and 3B).

We compared the *L. glauca* genome with the genomes of four Lauraceae plants, two magnoliids, one Juglandales plant, one Rosales plant, one Salicales plant, one Sapindales plant, one Euphorbiales plant, two Ranales plants, and one species of Amborellales (*Amborella trichopoda*) as the outgroups. Divergence estimates were based on fossil data for the node (*A. trichopoda*/*L. chinense* (173–199 Mya), *Cercidiphyllum japonicum*/*Prunus mume* (102–120 Mya), *L. cubeba*/*P. americana* (27–63 Mya), and *P. bournei*/*P. americana* (17–58 Mya)) and an ancient WGD event of *L. chinense* (116 Mya) (Chen et al., 2019). The phylogenetic tree based on 390 single-copy orthologs among the 15 species indicated that Piperales first diverged from the magnoliids at approximately 126.9–155.6 Mya, and Magnoliales (*L. chinense*) subsequently diverged from Laurales at approximately 119.8–135.6 Mya (Figure 3B). The divergence (the peak K_S differentiation ≈ 0.10) between *L. glauca* and *L. cubeba* was closer than that between other Lauraceae species analyzed; they diverged from their last common ancestor approximately 14.90–23.18 Mya (Figures 3A, 3B, and S10), with a substitution rate of 3.02E–9 mutations per site per year (Cui et al., 2006).

In angiosperms, gene family expansion or contraction facilitates evolutionary adaptation and trait innovations (Wang et al., 2019). The expansion or contraction of gene families can occur through WGDs or tandem duplications together with natural selection (Hartl and Clark, 2007). We estimated the ancestral gene content at each node of the species tree covering the 15 taxa, and significant changes along each branch were

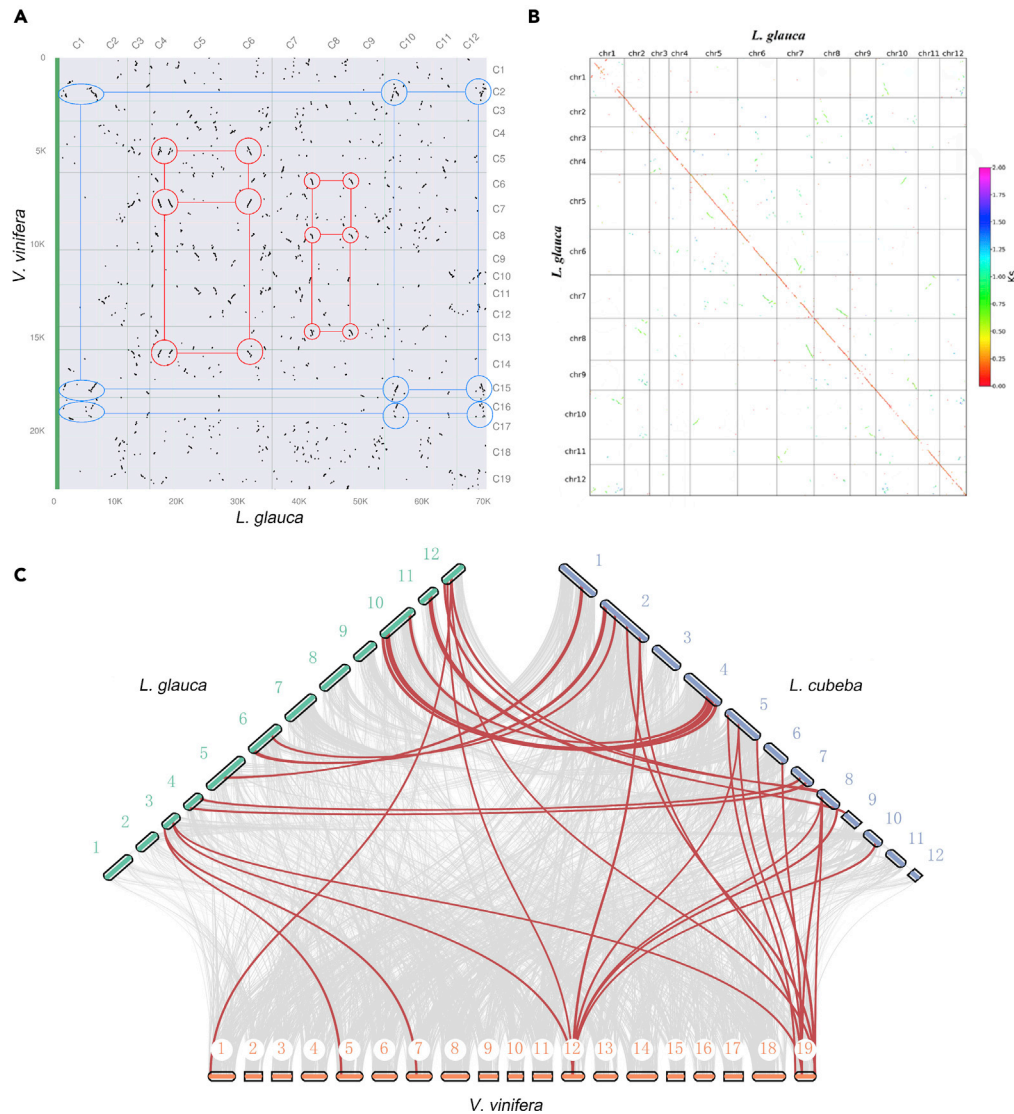


Figure 2. Intergenomic synteny

(A) Dot plots of orthologues between *L. glauca* and *V. vinifera*. The red and blue circles highlight examples of major duplication events which suggest 2:3 and 1:1 syntenic relationships, respectively.
 (B) Collinear point diagram of the *L. glauca* genome. Green copy region provides evidence of recent WGD and extra red copy (no diagonal) regions provide evidence of ancient WGD events.
 (C) Synteny patterns between genomic regions from two Lauraceae and *V. vinifera*. A colinear relationship (terpene synthase genes) is highlighted by one syntenic set shown in red colors.

modeled using CAFE software (Figure 3C). *L. glauca* exhibited the expansion of 1,544 gene families and contraction of 1,750 gene families compared with the inferred ancestral Lauraceae genome (Figure 3C, Tables S17, S18, S19, and S20). In the 20 genes displaying the most significant enrichment in the GO terms, eight, nine, and three of the expanded genes were significantly enriched in the pathways of molecular function, biological processes, and cellular component, respectively (Figure S4 and Table S19), whereas 2, 14, and four of the contracted genes were significantly enriched in the pathways of molecular function, biological process, and cellular component, respectively (Figure S7 and Table S17). Unique gene families in the *L. glauca* genome were enriched in the “interspecies interaction between organisms,” “anion transmembrane transporter activity,” and “plastid translation” GO functional categories, and in the “photo-transduction,” “aldosterone synthesis and secretion,” and “gastric acid secretion” KEGG pathways (Figures S2 and S3; Tables S10 and S11). The significantly expanded gene families were enriched in the

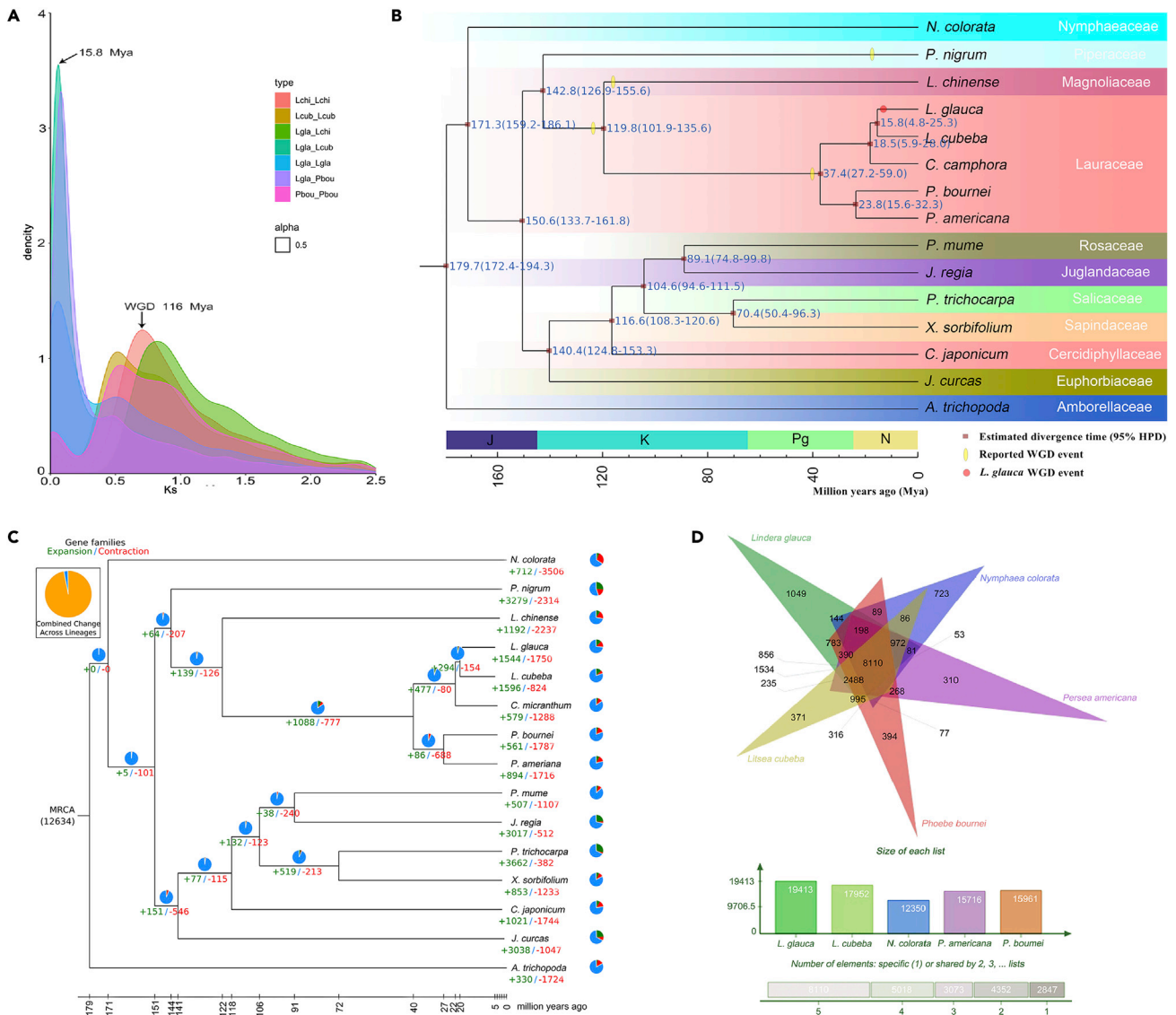


Figure 3. Comparative genomic analysis of *L. glauca* and WGD

(A) The distributions frequencies of synonymous substitution rate (Ks) for orthologs among four Lauraceae species.

(B) Phylogenetic tree with 390 single-copy orthologs from 15 species identified by OrthoMCL software to show divergence times (J Jurassic, K Cretaceous, Pg Paleogene, N Neogene).

(C) Gene family expansion/contraction identified by CAFE software. The green and red circles with corresponding numbers ($p < 0.05$) indicate gain (expansion) and loss (contraction) of gene families in different species, respectively.

(D) Sharing of gene families by *L. glauca* and four other Lauraceae species.

GO MRCA of ATPase activity, transmembrane transport, and auxin transport and in the KEGG pathways of ABC transporters and photosynthesis (Figures S4 and S5, Tables S12 and S13), whereas the contracted gene families were significantly enriched in the GO terms of the anchored component of membranes and protein heterodimerization activity but did not display significant enrichment in any KEGG pathway (Figures S6 and S7, Tables S14 and S15).

Evolution of *TPS* gene family

Terpenoids are synthesized from two direct precursor substrates—IPP and DMAPP—in two reactions: the first is catalyzed by FPP synthase (FPPS), GPP synthase, and GGPP synthase, and the second is catalyzed by TPSs (Nagegowda and Gupta, 2020). Thus, terpenoid production is associated with two major gene groups

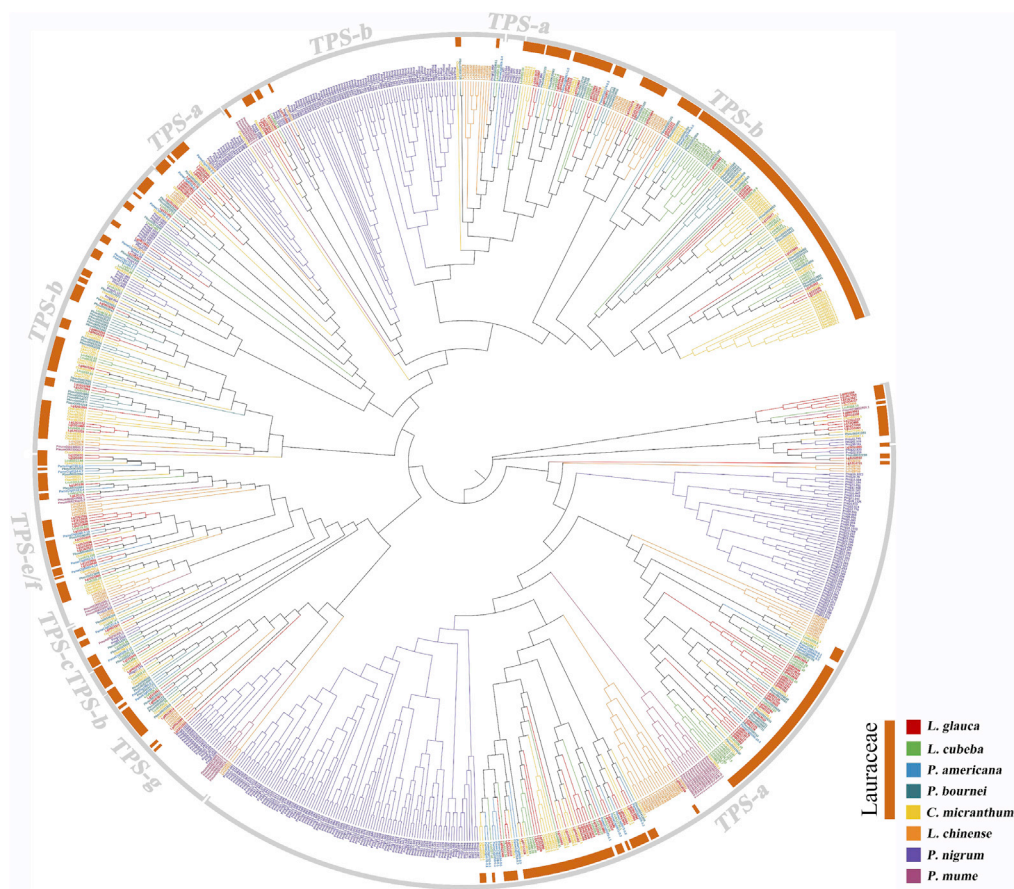


Figure 4. Phylogenetic placements of *L. glauca* TPS genes

The phylogenetic tree was constructed using putative or characterized TPS genes from eight sequenced plant genomes.

(Figure 5B): group I genes in the MVA pathway and group II genes in the MEP pathway (KEGG pathway: map00095). To determine the genomic basis of terpenoid biosynthesis, OrthoMCL was used to identify orthologous genes, and CAFE software was used to identify gene family clusters, focusing on expansion (gain) and contraction (loss) events related to the two gene groups in *L. glauca*. The results revealed 27,561 gene families comprising 285,460 genes across the 15 species (Figure 3D; for clarity, only Lauraceae species are shown). Thirteen gene families containing 1,118 genes were unique to *L. glauca*.

One of the most striking features of the *L. glauca* genome is many of TPS genes—138 LgTPS genes (Figure 4), which is the largest number for a genome of any Lauraceae species to date (Tables 2 and S21). These genes belong to six of seven gene subfamilies in gymnosperms and angiosperms (Chen et al., 2011), including TPS-a (69), TPS-b (45), TPS-c (1), TPS-e/f (15), and TPS-g (8). LgTPS genes in TPS-a, TPS-b, and TPS-g subfamilies encode proteins involved in specialized monoterpene (C10), sesquiterpene (C15), or diterpene (C20) biosynthesis. The genes in TPS-c (1) and TPS-e/f (15) subfamilies probably encode copalyl diterpene and kaurene synthases, which catalyze the formation of 20-carbon isoprenoids (C20s) and regulate plant primary metabolism (Martin et al., 2004; Chen et al., 2011). TPS-a and TPS-b subfamilies are the most diverse in the *L. glauca* genome, contributing to the mass and mixed production of volatile C15s and C10s (Sun et al., 2011).

LgTPS genes were not uniformly distributed across all chromosomes except for chromosome 1, and many members from individual gene subfamilies were clustered as tandem duplicates (Figure 5A). Chromosome 12 contained the most TPS genes (38). Specifically, 31 of them belonged to different subfamilies, namely LgTPS-a (10), LgTPS-b (20), and LgTPS-g (1), and were located in the 25–31-Mb region. Moreover, 107 LgTPS genes did not correspond to syntenic regions but were found elsewhere in the genome, and most of them belonged to the TPS-a subfamily.

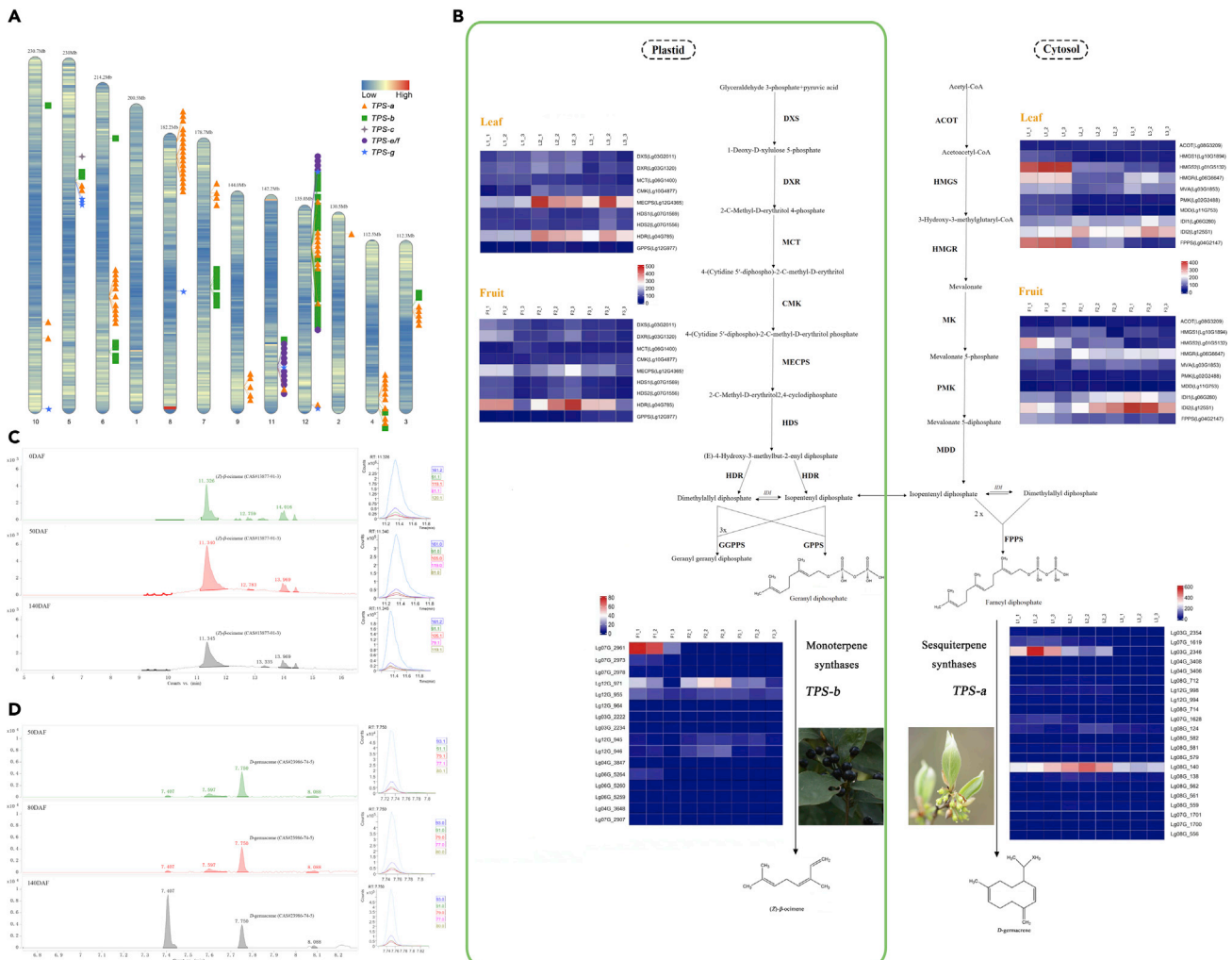


Figure 5. Scent biosynthesis of fruits and leaves in *L. glauca*

(A) in Distribution of LgTPS genes on chromosomes in *L. glauca*.

(B) MVA pathway mevalonate pathway and MEP pathway mevalonate-independent (deoxyxylulose phosphate) for the biosynthesis of (*Z*)- β -ocimene and *D*-germacrene and heatmap of related LgTPS genes. L1₋, L2₋, and L3₋ represent 20, 50, and 140 DAF of leaf. F1₋, F2₋, and F3₋ represent 50, 80, and 140 DAF of fruit. c-d Extracted ion chromatograms showing the content of the corresponding LgTPS gene expression products with three different stages of fruits and leaves.

Monoterpenes are generally formed from GPP by the action of monoterpene synthases (mTPSs), which belong to the *TPS-b*, *TPS-d-1*, and *TPS-g* subfamilies (Nagegowda, 2010; Ruan et al., 2016). In this study, 16 and two predicted genes displayed high homology with β -ocimene synthase and E- β -ocimene/myrcene synthase, which are involved in monoterpene (C₁₀H₁₆) synthesis. Phylogenetic analysis (Figure 4) revealed that these unigenes were clustered into *TPS-a* and *TPS-b* groups; we conjectured that *TPS-b* type monoterpene synthases in *L. glauca* fruit might be involved in specific monoterpene synthesis. A study reported that 13C₁₀H₁₆ monoterpene isomers are the main components of the essential oil of *L. glauca* fruit (Lin et al., 2017). β -ocimene synthase may be a multifunctional enzyme that can catalyze the synthesis of different monoterpenes, and their high sequence identity and functional diversity among predicted ocimene synthase members suggest the rapid evolution of a species-specific paralogous gene cluster in the *L. glauca* genome.

Comparative transcriptome analysis of terpenoid biosynthesis

In precursors, 412 and 231 unigenes in transcripts of *L. glauca* fruits and leaves were annotated to EMP and PPP pathways, respectively. A total of 286 and 126 unigenes were annotated to terpenoid biosynthesis and metabolism pathways, respectively, covering 15 enzymes involved in terpenoid synthesis pathways

(Figure 5B). Transcriptome sequencing of three stages (fruits at 50, 80, and 140 days after flowering [DAF]; leaves at 20, 50, and 140 DAF) identified 19,882 unigenes with differential expression. In total, 7,208 (4,311), 11,498 (6,772), and 5,344 (2,008) unigenes were specific for 50:80 (20:50) DAF, 50:140 (20:140), and 80:140 (50:140) DAF, respectively, for the fruits (leaves) differential transcripts.

The comparative transcriptome of *L. glauca* fruits and leaves revealed the upregulation and downregulation of catalytic synthase genes in the MVA (in the cytosol) and MEP (in plastid) pathways to the biosynthesis of terpenoids, respectively (Figure 5B). For monoterpene synthesis, most catalytic synthase genes had the highest expression in 80 DAF fruits and in 50 and 140 DAF leaves. *HDR* gene on chromosome 4 (*Lg04G785*) had signally expression in all three developmental stages, in both fruits and leaves. The expression level of *MECPS* on chromosome 12 (*Lg12G4365*) in leaves was significantly higher than that in fruits, especially in the last two stages. For sesquiterpene synthesis, although the expression trend of the catalytic synthase genes in fruits and leaves was similar, their expression levels in leaves were significantly higher than that in fruits. *HMG2* on chromosome 1 (*Lg01G5132*) and *FPPS* on chromosome 4 (*Lg04G2147*) in leaves had the highest expression at 20 DAF, with significantly higher expression than in fruits. The expression of two *IDI* genes, however, was the opposite. The expression level of *HMGR* on chromosome 6 (*Lg06G6647*) in leaves was the highest at 20 DAF, whereas in fruits, it was the highest at 140 DAF. Overall, the expression levels of genes for monoterpene synthesis were upregulated with fruit development, whereas those for sesquiterpene synthesis were downregulated with leaf development.

Identification of major single components and its scent biosynthesis-associated genes

The essential oils produced by *L. glauca* are widely used commercially and contain various compounds. GC-MS results revealed that the accumulation of terpenoids in fruits and leaves is at the early and middle-late stages (Figures 5C and 5D). Fruits at 140 DAF and leaves at 50 DAF exhibited the highest essential oil content of 0.78 mL/100 g and 0.23 mL/100 g, respectively, with a total of 74 and 25 constituents, respectively (Table S22). The monoterpene β -ocimene content was the highest (>30%) in 140 DAF fruits, and *D*-germacrene content was the highest in 140 DAF leaves (>40%). The result was consistent with patterns of essential oil content and changes in essential oil composition during fruit development and is supported by some studies (Sun et al., 2011; Zhu et al., 2016).

TPSs are the rate-limiting enzymes in the production of plant terpenoids (Tholl, 2006). *L. glauca* has numerous TPS-a and TPS-b members, which regulate the biosynthesis of monoterpenes and sesquiterpenes. Comparative transcriptome analysis revealed the expression levels of TPS-a and TPS-b members in *D*-germacrene (sesquiterpene) and β -ocimene (monoterpene) synthesis of *L. glauca* (Figure 5B). A total of 16 TPS-a and 22 TPS-b genes were significantly expressed during the leaf and fruit development stages, respectively. *Lg07G2961* on chromosome seven and *Lg12G971* on chromosome 12 shared the most considerable accumulation of major aromatic compounds during fruit development. Moreover, *Lg03G2346* on chromosome three and *Lg08G140* on chromosome 12 shared the most considerable accumulation of major aromatic compounds during leaf development. In addition, *Lg07G2961* expressions reached their peak at the 50 DAF of fruit development and then began to rapidly decrease, whereas *Lg12G971* had the highest expression at 80 DAF. *Lg3G2346* was significantly expressed at 20 DAF of leaf development and then decreased gradually. *Lg08G140* expression was stable in all stages of leaf development, and the peak was at 50 DAF. In summary, genes related to terpenoid biosynthesis were highly expressed in the early stage of fruit and leaf development, consistent with the significant accumulation of essential oil at this stage.

DISCUSSION

The Lauraceae family is well-suited for studying the linkage between phytochemistry, ecology, and evolution because of its characteristic secondary metabolism (Chen et al., 2011; Chaw et al., 2019), especially given its diversity of aromatic compounds (Chen et al., 2020c). The *L. glauca* genome is a valuable resource for phylogenomic analyses and studies of the TPS family, as we demonstrated here by elucidating mechanisms underlying the synthesis of its terpenoid products. The chromosome-level genome assembly also enhanced the status of *L. glauca* as a model in the *Lindera* genus. Our findings are expected to further the understanding of the biosynthesis of some natural products that are essential in traditional Chinese medicines, such as flavonoids (Huh et al., 2014) and lignan (Kim et al., 2014).

A final set of 65,145 protein-coding genes were predicted in the *L. glauca* genome, considerably higher than that predicted in other species in the Lauraceae family (Table S7), including *Cinnamomum camphora*

Table 2. Numbers of TPS subfamilies in 13 sequenced plant genomes

TPS subfamilies		a	b	c	d ^a b	e/f	g	
Function species	Genome size ^a (Mb)	C15	C10 (IspS)	C20 (CPS)	C10, C15, C20	C20	C10	Total number
Gymnosperms								
<i>G. biloba</i>	10,609	–	–	1	49	1	–	51
<i>P. abies</i>	12,301	–	–	2	59	1	–	62
Angiosperms								
<i>A. thaliana</i>	120	23	13	1	–	2	1	32
<i>A. trichopoda</i>	706	–	7	1	–	4	5	17
<i>L. chinense</i>	1,742	34	27	3	–	10	8	82
<i>V. vinifera</i>	434	29	10	2	–	3	14	58
<i>P. mume</i>	237	13	6	1	–	5	6	31
<i>P. nigrum</i>	761	192	81	2	–	1	17	293
Lauraceae								
<i>C. kanehirae</i>	731	25	58	2	–	12	4	101
<i>P. americana</i> ^{^c}	912	18	6	2	–	7	2	35
<i>P. bournei</i>	989	17	44	3	–	7	13	84
<i>L. cubeba</i>	1,370	28	44	2	–	7	7	88
<i>L. glauca</i>	2,092	69	45	1	–	15	8	138

Note: a The genome sizes of all species are based on published data. b TPS-d subfamily is gymnosperm specific. c *P. americana*, Hass cultivar (Rendón-Anaya et al., 2019). IspS, isoprene synthase; CPS, copalyl diphosphate synthase.

(27,899; Chaw et al., 2019), *L. cubeba* (31,329; Chen et al., 2020a), *P. bournei* (28,198; Chen et al., 2020b), and *P. americana* (Hass cultivar; 24,616; Rendón-Anaya et al., 2019). Most of these *L. glauca* genes were not grouped in homologous gene clusters (Figure S11) and were enriched in the GO terms of the recognition of pollen, UDP-glycosyltransferase activity, and fatty acid biosynthetic process (Figure S12) and in the KEGG pathways of glutathione metabolism, purine metabolism, and glycolysis (Figure S13).

Ancient WGD (also known as polyploidization) events are crucial driving forces in the evolution of fauna, flora, and fungi (Jiao et al., 2011; Adams, 2013). In the present study, K_S analysis indicated that the most recent WGD events of the *L. glauca* lineage closely followed the divergence of many Lauraceae genera, perhaps being the key contributor to Lauraceae diversification. The K_S peak value for the recent WGD event in *L. glauca* was smaller than the orthologous K_S peak value between *L. glauca* and *L. cubeba*, suggesting that the recent WGD event had occurred shortly after the divergence of Lauraceae, and this may have caused *L. glauca* to have a considerably larger genome and more protein-coding genes than *L. cubeba* and other Lauraceae species. In addition, the other Lauraceae species experienced two ancient WGD events before the divergence of the common ancestors of Lauraceae, implying that *L. glauca* experienced a recent WGD event after the groups diverged, thus explaining the abundance of gene models. That is similar to the *Piper nigrum* genome with 761.74 Mb, containing 63,466 genes and most of them unclustered.

Consistent with the results of early isozyme analysis and genome studies (Soltis and Soltis, 1990), we identified two independent WGD events before Lauraceae species diverged, and one recent WGD event of *L. glauca* genome, which may have contributed to gene family expansions and innovations in the interaction of different organisms and generated numerous TPS genes. Gene tree topologies (Figure 4) for each of the six angiosperm TPS subfamilies revealed the diversification of TPS genes and gene function in the ancestry of *L. glauca*. Our results (Table 2) indicated massive diversification of TPS-a and TPS-b subfamilies within the Lauraceae family, which is consistent with previous reports; the genes of these subfamilies encode the main constituents of 58 essential oils produced in *Cinnamomum* leaves, namely monoterpene (C10s) and sesquiterpene (C15s) (Cheng et al., 2015; Chaw et al., 2019), which are the main aroma constituents of fruits and leaves of *L. cubeba* (Si et al., 2012; Chen et al., 2020c) and *L. glauca* (Sun et al., 2011; Zhu et al., 2016). The number of C20s products of Lauraceae species, encoded by TPS-e/f genes, exceeded those of other angiosperms and may be Lauraceae-specific because this plant group contains numerous terpene compounds

(Chen et al., 2011). In our study, candidate TPSs genes were identified in the KEGG database using BLASTP (v.2.6.0), which may have led to overidentification. However, *L. chinense*, *P. mume*, *P. nigrum*, and four other Lauraceae species were predicted for TPSs genes by using the same method (Table 2), suggesting that *L. glauca* still had the largest TPS family among the Lauraceae species published.

Similar to other angiosperms, the initial precursors in terpenoid biosynthesis of *L. glauca* are acetyl-CoA, pyruvate, and D-glyceraldehyde-3-phosphate, which are produced in the PPP and EMP pathways (Tholl, 2006). Our analyses of the critical enzymes involved in the monoterpene and sesquiterpene biosynthesis pathways suggest that the overexpression of the related genes (*Lg07G2961* and *Lg12G971* in *TPS-b*; *Lg03G2346* and *Lg08G140* in *TPS-a*) may enhance the production of the principal components in terpene, *D*-germacrene, and β -ocimene (Figure 5). Terpenoid accumulation in *L. glauca* is probably owing to TPS family size and regulation of gene expression. The variations of terpene production and concentration among individual species are mainly the result of gene regulation divergence, functional differentiation, and cluster formation in the TPS family (Chen et al., 2020a). In addition, considering a severe male/female ratio imbalance in wild populations throughout East Asia (Xiong et al., 2018), unlike in Lauraceae species (Dupont, 2002), the newly replicated TPS genes in *L. glauca* may be related to the newly evolved terpenoids that protect plants from biological and abiotic stresses (Pichersky and Raguso, 2018).

In conclusion, the current assembly of the *L. glauca* genome and the analysis of Lauraceae species with available genome sequences provide insights into the underlying genetic diversity following WGD events responsible for the expansion and contraction of the TPS gene family. We analyzed the biological processes related to major single components of terpenoid biosynthesis in *L. glauca* fruits and leaves and identified four associated *TPS-b* and *TPS-a* genes.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
 - Lead contact
 - Materials availability
 - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
 - Plant materials and genome sequencing
 - Genome assembly and assessment
 - Hi-C library construction and chromosome assembly
 - Genome annotation
 - Comparative genomics for phylogenomic analysis and gene family
 - Whole-genome duplication and colinearity analyses
 - TPS gene identification
 - Metabolomic and transcriptome analysis
- QUANTIFICATION AND STATISTICAL ANALYSIS
- ADDITIONAL RESOURCES

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at <https://doi.org/10.1016/j.isci.2022.104761>.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (31900272; 31972948), Guizhou Science and Technology Plan Project (Qiankehe Basics-ZK [2021] General 151), Guizhou Science and Technology Support Project (Qiankehe Support [2020]1Y146), College and Education committee project of Guizhou University (703/702534183301), and Cultivation Project of Guizhou University (Gzu.2020 No.65). We thank Andrew H Paterson (University of Georgia) for suggestions and review of the entire article.

AUTHOR CONTRIBUTIONS

ZZ, SD, and BX conceived the project and designed the study; BX and LM performed the sampling; BX, LX, SL, TZ, and LL performed the experiments; BX, XH, and YN performed the data analysis; BX and LM designed and visualized the figures; BX and SD wrote the article; and all authors read and approved the final article.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: December 28, 2021

Revised: April 27, 2022

Accepted: July 10, 2022

Published: August 19, 2022

REFERENCES

- Adams, K. (2013). Genomic clues to the ancestral flowering plant. *Science* 342, 1456–1457.
- Bulotta, S., Corradino, R., Celano, M., D'Agostino, M., Maiuolo, J., Oliverio, M., Procopio, A., Iannone, M., Rotiroli, D., and Russo, D. (2011). Antiproliferative and antioxidant effects on breast cancer cells of oleuropein and its semisynthetic peracetylated derivatives. *Food Chem.* 127, 1609–1614.
- Burton, J.N., Adey, A., Patwardhan, R.P., Qiu, R., Kitzman, J.O., and Shendure, J. (2013). Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat. Biotechnol.* 31, 1119–1125.
- Chaw, S.M., Liu, Y.C., Wu, Y.W., Wang, H.Y., Lin, C.Y.I., Wu, C.S., Ke, H.M., Chang, L.Y., Hsu, C.Y., Yang, H.T., et al. (2019). Stout camphor tree genome fills gaps in understanding of flowering plant genome evolution. *Nat. Plants* 5, 63–73.
- Chen, F., Tholl, D., Bohlmann, J., and Pichersky, E. (2011). The family of terpene synthases in plants: a mid-size family of genes for specialized metabolism that is highly diversified throughout the kingdom. *Plant J.* 66, 212–229.
- Chen, W., Gong, L., Guo, Z., Wang, W., Zhang, H., Liu, X., Yu, S., Xiong, L., and Luo, J. (2013). A novel integrated method for large-scale detection, identification, and quantification of widely Targeted metabolites: application in the study of rice metabolomics. *Mol. Plant* 6, 1769–1780.
- Chen, J., Hao, Z., Guang, X., Zhao, C., Wang, P., Xue, L., Zhu, Q., Yang, L., Sheng, Y., Zhou, Y., et al. (2019). Liriodendron genome sheds light on angiosperm phylogeny and species-pair differentiation. *Nat. Plants* 5, 18–25.
- Chen, H., Köllner, T.G., Li, G., Wei, G., Chen, X., Zeng, D., Qian, Q., and Chen, F. (2020a). Combinatorial evolution of a terpene synthase gene cluster explains terpene variations in *Oryza*. *Plant Physiol.* 182, 480–492.
- Chen, S.P., Sun, W.H., Xiong, Y.F., Jiang, Y.T., Liu, X.D., Liao, X.Y., Zhang, D.Y., Jiang, S.Z., Li, Y., Liu, B., et al. (2020b). The Phoebe genome sheds light on the evolution of magnoliids. *Hortic. Res.* 7, 146.
- Chen, Y.C., Li, Z., Zhao, Y.X., Gao, M., Wang, J.Y., Liu, K.W., Wang, X., Wu, L.W., Jiao, Y.L., Xu, Z.L., et al. (2020c). The Litsea genome and the evolution of the laurel family. *Nat. Commun.* 11, 1675.
- Cheng, S.S., Lin, C.Y., Yang, C.K., Chen, Y.J., Chung, M.J., and Chang, S.T. (2015). Chemical polymorphism and composition of leaf essential oils of *Cinnamomum micranthum* using gas chromatography/mass spectrometry, cluster analysis, and principal component analysis. *J. Wood Chem. Technol.* 35, 207–219.
- Christianson, D.W. (2017). Structural and chemical biology of terpenoid cyclases. *Chem. Rev.* 117, 11570–11648.
- Cordoba, E., Salmi, M., and León, P. (2009). Unravelling the regulatory mechanisms that modulate the MEP pathway in higher plants. *J. Exp. Bot.* 60, 2933–2943.
- Cui, L., Wall, P.K., Leebens-Mack, J.H., Lindsay, B.G., Soltis, D.E., Doyle, J.J., Soltis, P.S., Carlson, J.E., Arumuganathan, K., Barakat, A., et al. (2006). Widespread genome duplications throughout the history of flowering plants. *Genome Res.* 16, 738–749.
- De Bie, T., Cristianini, N., Demuth, J.P., and Hahn, M.W. (2006). CAFE: a computational tool for the study of gene family evolution. *Bioinformatics* 22, 1269–1271.
- Dupont, Y.L. (2002). Evolution of apomixis as a strategy of colonization in the dioecious species *Lindera glauca* (Lauraceae). *Popul. Ecol.* 44, 293–297.
- Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797.
- Enfissi, E.M.A., Fraser, P.D., Lois, L.M., Boronati, A., Schuch, W., and Bramley, P.M. (2005). Metabolic engineering of the mevalonate and non-mevalonate isopentenyl diphosphate-forming pathways for the production of health-promoting isoprenoids in tomato. *Plant Biotechnol. J.* 3, 17–27.
- Gremme, G., Steinbiss, S., and Kurtz, S. (2013). GenomeTools: a comprehensive software library for efficient processing of structured genome annotations. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 10, 645–656.
- Gutensohn, M., Nagegowda, D.A., and Dudareva, N. (2013). Involvement of compartmentalization in monoterpene and sesquiterpene biosynthesis in plants. *Springer Sci.* 5, 155–169.
- Hartl, D.L., and Clark, A.G. (2007). Principles of population genetics, 7 (*Ecoscience*), p. 372.
- Hoff, K.J., Lange, S., Lomsadze, A., Borodovsky, M., and Stanke, M. (2015). BRAKER1: unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics* 32, 767–769.
- Hu, L., Xu, Z., Wang, M., Fan, R., Yuan, D., Wu, B., Wu, H., Qin, X., Yan, L., Tan, L., et al. (2019). The chromosome-scale reference genome of black pepper provides insight into piperine biosynthesis. *Nat. Commun.* 10, 4702.
- Huh, G.W., Park, J.H., Kang, J.H., Jeong, T.S., Kang, H.C., and Baek, N.I. (2014). Flavonoids from *Lindera glauca* Blume as low-density lipoprotein oxidation inhibitors. *Nat. Prod. Res.* 28, 831–834.
- Jiao, Y., Wickett, N.J., Ayyampalayam, S., Chandrabali, A.S., Landherr, L., Ralph, P.E., Tomsho, L.P., Hu, Y., Liang, H., Soltis, P.S., et al. (2011). Ancestral polyploidy in seed plants and angiosperms. *Nature* 473, 97–100.
- Kim, K.H., Moon, E., Ha, S.K., Suh, W.S., Kim, H.K., Kim, S.Y., Choi, S.U., and Lee, K.R. (2014). Bioactive lignan constituents from the twigs of *Lindera glauca*. *Chem. Pharm. Bull.* 62, 1136–1140.
- Kim, Y.S., Kim, E.K., Dong, X., Park, J.S., Shin, W.B., Kim, S.J., Go, E.A., Park, P.J., and Lim, B.O. (2019). *Lindera glauca* (Siebold et Zucc.) Blume Stem Extracts Protect Against tert-Butyl Hydroperoxide-Induced Oxidative Stress. *J. Med. Food* 22, 508–520.
- Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H., and Phillippy, A.M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* 27, 722–736.
- Kramer, E.M., and Irish, V.F. (1999). Evolution of genetic mechanisms controlling petal development. *Nature* 399, 144–148.

- Li, H., and Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature* 475, 493–496.
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The sequence alignment/map (SAM) format and SAMtools. *Bioinformatics* 25, 2078–2079.
- Liao, P., Wang, H., Hemmerlin, A., Nagegowda, D.A., Bach, T.J., Wang, M., and Chye, M.L. (2014). Past achievements, current status and future perspectives of studies on 3-hydroxy-3-methylglutaryl-CoA synthase (HMGS) in the mevalonate (MVA) pathway. *Plant Cell Rep.* 33, 1005–1022.
- Lin, Z., An, J., Wang, J., Niu, J., Ma, C., Wang, L., Yuan, G., Shi, L., Liu, L., Zhang, J., et al. (2017). Integrated analysis of 454 and Illumina transcriptomic sequencing characterizes carbon flux and energy source for fatty acid synthesis in developing *Lindera glauca* fruits for woody biodiesel. *Biotechnol. Biofuels* 10, 134.
- Liu, H., Wu, S., Li, A., and Ruan, J. (2021). SMARTdenovo: a de novo assembler using long noisy reads. *Gigabyte* 2021, 1–9.
- Luo, J., Lyu, M., Chen, R., Zhang, X., Luo, H., and Yan, C. (2019). SLR: a scaffolding algorithm based on long reads and contig classification. *BMC Bioinf.* 20, 539.
- Marçais, G., and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* 27, 764–770.
- Martin, D.M., Fäldt, J., and Bohlmann, J. (2004). Functional characterization of nine Norway spruce *TPS* genes and evolution of gymnosperm terpene synthases of the *TPS-d* subfamily. *Plant Physiol.* 135, 1908–1927.
- Martin, D.M., Aubourg, S., Schouwey, M.B., Daviet, L., Schalk, M., Toub, O., Lund, S.T., and Bohlmann, J. (2010). Functional annotation, genome organization and phylogeny of the grapevine (*Vitis vinifera*) terpene synthase gene family based on genome assembly, FLDNA cloning, and enzyme assays. *BMC Plant Biol.* 10, 226.
- Nagegowda, D.A., and Gupta, P. (2020). Advances in biosynthesis, regulation, and metabolic engineering of plant specialized terpenoids. *Plant Sci.* 294.
- Nagegowda, D.A. (2010). Plant volatile terpenoid metabolism: biosynthetic genes, transcriptional regulation and subcellular compartmentation. *FEBS Lett.* 584, 2965–2973.
- Nawrocki, E.P., and Eddy, S.R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* 29, 2933–2935.
- Park, S., Song, J.H., Nhiem, N.X., Ko, H.J., and Kim, S.H. (2018). The chemical constituents from twigs of *Lindera glauca* (Siebold & Zucc.) Blume and their antiviral activities. *Phytochem. Lett.* 25, 74–80.
- Pichersky, E., and Raguso, R.A. (2018). Why do plants produce so many terpenoid compounds. *New Phytol.* 220, 692–702.
- Pu, X., Li, Z., Tian, Y., Gao, R., Hao, L., Hu, Y., He, C., Sun, W., Xu, M., Peters, R.J., et al. (2020). The honeysuckle genome provides insight into the molecular mechanism of carotenoid metabolism underlying dynamic flower coloration. *New Phytol.* 227, 930–943.
- Qi, J., Xiong, B., Zhang, Z.X., Zhen, A.G., and Pu, F.G. (2016). Diversity analysis on fatty acid composition of *Lindera glauca* fruit from different provenances. *Journal of the Chinese Cereals and Oils Association* 31, 72–78.
- Rendón-Anaya, M., Ibarra-Laclette, E., Méndez-Bravo, A., Lane, T., Zheng, C., and Carretero-Pauletg, L. (2019). The avocado genome informs deep angiosperm phylogeny, highlights introgressive hybridization, and reveals pathogen-influenced gene space adaptation. *Proc. Natl. Acad. Sci. USA* 116, 17081–17089.
- Rohdich, F., Zepeck, F., Adam, P., Hecht, S., Kaiser, J., Laupitz, R., Gräwert, T., Amslinger, S., Eisenreich, W., Bacher, A., and Arigoni, D. (2003). The deoxyxylulose phosphate pathway of isoprenoid biosynthesis: studies on the mechanisms of the reactions catalyzed by IspG and IspH protein. *Proc. Natl. Acad. Sci. USA* 100, 1586–1591.
- Ruan, J.X., Li, J.X., Fang, X., Wang, L.J., Hu, W.L., Chen, X.Y., and Yang, C.Q. (2016). Isolation and characterization of three new monoterpene synthases from *Artemisia annua*. *Front. Plant Sci.* 7, 638.
- Ruan, Q.F., Jiang, S.Q., Zheng, X.Y., Tang, Y.Q., Yang, B., Yi, T., Jin, J., Cui, H., and Zhao, Z. (2020). Pseudoguaianolactones A–C: three unusual sesquiterpenoids from *Lindera glauca* with anti-inflammatory activities by inhibiting the LPS-induced expression of iNOS and COX-2. *Chem. Commun.* 56, 1517–1520.
- Servant, N., Varoquaux, N., Lajoie, B.R., Viara, E., Chen, C.J., Vert, J.P., Heard, E., Dekker, J., and Barillot, E. (2015). HiC-Pro: an optimized and flexible pipeline for Hi-C processing. *Genome Biol.* 16, 259.
- Si, L., Chen, Y., Han, X., Zhan, Z., Tian, S., Cui, Q., and Wang, Y. (2012). Chemical composition of essential oils of *Litsea cubeba* harvested from its distribution areas in China. *Molecules* 17, 7057–7066.
- Soltis, D.E., and Soltis, P.S. (1990). Isozyme evidence for ancient polyploidy in primitive angiosperms. *Syst. Bot.* 15, 328–337.
- Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 1312–1313.
- Sun, H.L., Wang, J.X., Gu, X.Z., and Kang, W.Y. (2011). Analysis of volatile components from leaves and fruits of *Lindera glauca*. *Chinese J. Exp. Prescrip.* 17, 9497.
- Suh, W.S., Kim, K.H., Kim, H.K., Choi, S.U., and Lee, K.R. (2015). Three new lignan derivatives from *Lindera glauca* (Siebold et Zucc.) Blume. *Helvetica Chimica Acta* 98, 1087–1094. <https://doi.org/10.1002/hlca.201500002>.
- Sun, P., Jiao, B., Yang, Y., Shan, L., Li, T., Li, X., Xi, Z., Wang, X., and Liu, J. (2021). WGD: a user-friendly toolkit for evolutionary analyses of whole-genome duplications and ancestral karyotypes. Preprint at bioRxiv. <https://doi.org/10.1101/2021.04.29.441969>.
- Tholl, D. (2006). Terpene synthases and the regulation, diversity and biological roles of terpene metabolism. *Curr. Opin. Plant Biol.* 9, 297–304.
- Tsui, H.P., and Werff, H. (2008). Lauraceae and Hernandiaceae. In *Flora of China*, 7, H.W. Li, ed (Beijing & St. Louis: Science Press & Missouri Botanical Garden Press), pp. 146–147.
- Vaser, R., Sović, I., Nagarajan, N., and Šikić, M. (2017). Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* 27, 737–746.
- Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C.A., Zeng, Q., Wortman, J., Young, S.K., and Earl, A.M. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* 9, e112963.
- Wang, W., Cheng, Y., Chen, D., Liu, D., Hu, M., Dong, J., Zhang, X., Song, L., and Shen, F. (2019). The Catalase Gene Family in Cotton: Genome-Wide Characterization and Bioinformatics Analysis. *Cells* 8, 86. <https://doi.org/10.3390/cells8020086>.
- Wang, Y., Gao, X., Yu, X., Cheng, S., and Kong, L. (1994). Study on the resource and its utilizations of *Lindera glauca* in China. *Henan Science* 12, 331–334.
- Wang, Y., Tang, H., DeBarry, J.D., Tan, X., Li, J., Wang, X., Lee, T.H., Jin, H., Marler, B., Guo, H., et al. (2012). MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and colinearity. *Nucleic Acids Res.* 40, e49.
- Xiong, B., Dong, S.B., and Zhang, Z.X. (2018). Biodiesel from *Lindera glauca* oil, a potential non-food feedstock in Southern China. *Industrial Crops & Products* 122, 107–113. <https://doi.org/10.1016/j.indcrop.2018.05.026>.
- Yang, Z. (2007). Paml 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24, 1586–1591.
- Zhu, B., Hou, X., Niu, J., Li, P., Fang, C., Qiu, L., Ha, D., Zhang, Z., Sun, J., Li, Y., and Lin, S. (2016). Volatile Constituents from the Fruits of *Lindera glauca* (Sieb. et Zucc.) with Different Maturities. *J. Essent. Oil Bear. Plants* 19, 926–935.
- Zulak, K.G., Lippert, D.N., Kuzyk, M.A., Domanski, D., Chou, T., Borchers, C.H., and Bohlmann, J. (2009). Targeted proteomics using selected reaction monitoring reveals the induction of specific terpene synthases in a multi-level study of methyl jasmonate-treated Norway spruce (*Picea abies*). *Plant J.* 60, 1015–1030.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological samples		
The total DNA and RNA of <i>Lindera glauca</i> leaves and fruits	This study	NA
Critical commercial assays		
DNAsecure Plant Kit	TIANGEN	DP320-03
RNAprep Pure Plant Kit	TIANGEN	DP432
Deposited data		
Assembly genome	This paper	GenBank: PRJNA733927
Software and algorithms		
Jellyfish	Marçais and Kingsford, 2011	http://www.genome.umd.edu/jellyfish.html
CANU v.1.7	Koren et al., 2017	https://github.com/marbl/canu
SMARTdenovo	NA	https://github.com/ruanjue/smartdenovo
RACON v.1.4.12	Vaser et al., 2017	https://github.com/lbcb-sci/racon
PILON v.1.22	Walker et al., 2014	https://github.com/broadinstitute/pilon
BUSCO v.4.1.2	NA	https://busco.ezlab.org/
Hi-C Pro	Servant et al., 2015	https://github.com/nservant/HiC-Pro
SAMtools	Li et al., 2009	https://github.com/samtools/samtools
LACHESIS	Burton et al., 2013	https://github.com/shendurelab/LACHESIS
SLR	Luo et al., 2019	https://github.com/luojunwei/SLR
RepeatModelers	NA	http://www.repeatmasker.org/RepeatModeler/
GenomeTools	Gremme et al., 2013	https://github.com/genometools/genometools
TransposonPSI	NA	http://transposonpsi.sourceforge.net/
RepeatMasker v.4.0.5	NA	http://www.repeatmasker.org/cgi-bin/RepeatProteinMaskRequest
BRAKER	Hoff et al., 2015	https://github.com/Gaius-Augustus/BRAKER
Rfam	Nawrocki and Eddy, 2013	https://github.com/thaines/rfam
BLASTP	NA	https://blast.ncbi.nlm.nih.gov/Blast.cgi
OrthoMCL v.2.0.9	Kramer and Irish, 1999	https://github.com/stajichlab/OrthoMCL
MUSCLE v.3.8.31	Edgar, 2004	https://github.com/rcedgar/muscle
TRIMAL v.1.4.rev22	NA	http://trimal.cgenomics.org/introduction
RAxML v.8.2.10	Stamatakis, 2014	https://github.com/stamatak/standard-RAxML
CAFE v4.2	De Bie et al., 2006	https://github.com/hahnlab/CAFE
MCSanX	Wang et al., 2012	https://github.com/wyp1125/MCSanX
PAML v.4.9	Yang, 2007	https://github.com/abacus-gene/paml
WGDI	Sun et al., 2021	https://github.com/SunPengChuan/wgdi
Other		
Nanopore sequencing	Oxford Nanopore Technologies	Oxford Nanopore platform
Illumina sequencing	Illumina Technologies	Illumina NovaSeq 6000 / HiSeq X platform
GC-MS analysis	Santa Clara, CA, USA	Agilent 5975B

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Biao Xiong (bxiong@gzu.edu.cn).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- The accession number for the genome assembly and raw reads reported in this paper is GenBank: PRJNA733927 with BioSample accessions SAMN19468007, including Illumina RNA-Seq data.
- This paper does not report original code.
- Any additional information required to reanalyze the data reported in this paper is available from the [lead contact](#) upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

This study does not include experiments model or subjects.

METHOD DETAILS

Plant materials and genome sequencing

A wild-type elite female *L. glauca* plant ($2n = 24$), which was transplanted from a nursery (31.8°N, 114.1°E) established at Jigongshan National Nature Reserve (Henan, China) to the laboratory, was used for sequencing (Supplementary 1). Genomic DNA and total RNA were extracted from fresh leaves by using the DNaseq Plant Kit and RNAPrep Pure Plant Kit (TIANGEN, Germany), respectively. For genome survey analysis and genome polishing, a short paired-end Illumina DNA library with a 350-bp insert size was sequenced on the NovaSeq 6000/HiSeq X Ten platform, according to the manufacturer's instructions (Illumina, San Diego, CA, USA). For long-read DNA sequencing, high-quality genomic DNA fragments (>20 kb) were used to construct long-read libraries on the ONT platform (<https://nanoporetech.com>) and were sequenced using ONT long reads.

Genome assembly and assessment

After the genome size was estimated, the heterozygosity and repeat rate of *L. glauca* were determined using the K-mer method with Jellyfish software (<http://www.genome.umd.edu/jellyfish.html>), and the assembly of the genome was sequenced using ONT long reads and 10× Genomics linked-reads technology. The ONT reads were corrected and trimmed using CANU (v.1.7) (Koren et al., 2017). The corrected ONT reads were directly assembled using SMARTdenovo (<https://github.com/ruanjue/smartdenovo>) (Liu et al., 2021). The contigs for assembly were polished twice with the Illumina short reads by using the RACON (v.1.4.12) (Vaser et al., 2017) and PILON (v.1.22) (Walker et al., 2014) software packages. To evaluate the consistency and completeness of the genome assembly, multiply analyses were performed, including the determination of the map rate and average depth of the Illumina short reads, BUSCO (v.4.1.2) alignment, core eukaryotic genes alignment, and RNA-Seq read mapping.

Hi-C library construction and chromosome assembly

For Hi-C library construction, *L. glauca* seedlings were fixed with 2% formaldehyde solution at room temperature for 40 min in vacuum and then fixed for the crosslinking reaction. The fixed tissue was ground with liquid nitrogen for the primary library preparation. Thereafter, extracted nuclei were resuspended in 0.5% Sodium dodecyl sulfate (SDS) and incubated at 60 °C for 10 min, and the SDS was subsequently quenched with 10% Triton X-100 (Chen et al., 2020b). The DNA was digested with the DpnII restriction enzyme.

The cohesive DNA ends were filled and marked with biotin-labeled dCTP and dCAP, dTTP, and dGTP by using Klenow, and the filled-in DpnII sites were ligated to construct proximal chromatin DNA. After these complexes were purified and sheared, biotinylated Hi-C ligation products were extracted and used to construct Illumina sequencing libraries, according to the standard Illumina library preparation protocol. The libraries were sequenced on HiSeq X Ten DNA sequencers to obtain PE150 reads, and Hi-C Pro

software was used to validate these reads (Servant et al., 2015). The high-quality reads were mapped to the draft scaffolds based on fast and accurate short read alignment with Burrows–Wheeler transform (Li and Durbin, 2011), and unmapped reads were removed using SAMtools (Li et al., 2009). The Hi-C read pairs mapped in draft scaffolds were clustered and ordered into chromosomes in accordance with chromatin interactions using LACHESIS software (Burton et al., 2013). Subsequently, the contigs were independently assembled into scaffolds using the SLR program (Luo et al., 2019).

Genome annotation

The combination of the *de novo* repeat library and homology-based strategies were used to identify repeat structures and trace repeat elements in the *L. glauca* genome by using RepeatModelers (<http://www.repeatmasker.org/RepeatModeler/>), the GenomeTools suite (Gremme et al., 2013), and TransposonPSI (<http://transposonpsi.sourceforge.net/>). Subsequently, RepeatMasker software (v.4.0.5) was used to annotate repeat sequences. Coding genes were predicted using the MAKER (v.2.31.8) pipeline (softmask = 1, augustus_species = tomato, est2genome = 1, protein2genome = 1), integrating transcript sequences from the *de novo* assembly of the *L. glauca* Illumina RNA-seq data. BRAKER (Hoff et al., 2015) (–softmasking, –downsampling_lambda 1) was performed for gene structure annotations; it integrates GeneMark-ET and AUGUSTUS by combining the aligned results. Moreover, noncoding RNAs, small RNAs, and non-protein-coding genes were annotated by aligning to Rfam using INFERNAL (Nawrocki and Eddy, 2013) and performing homologous searching and deep-learning analyses across the assembled genome sequence. BUSCO software (version: 4.1.2) was used to examine the completeness of genomic annotations.

Comparative genomics for phylogenomic analysis and gene family

To investigate the evolutionary history of the *L. glauca* genome, the relationships of its protein-coding genes with those of 14 other angiosperms (*L. cubeba*, *Cinnamomum camphora*, *Phoebe bournei*, *Persea americana*, *Piper nigrum*, *Armeniaca mume*, *Cercidiphyllum japonicum*, *Jatropha curcas*, *Xanthoceras sorbifolium*, *Aquilegia viridiflora*, *Populus trichocarpa*, *Liriodendron chinense*, *Nymphaea colorate*, *Amborella trichopoda*, and *Juglans regia*) were analyzed. All amino acid sequences of selected plants were compared using BLASTP (v.2.6.0) (*E-value* cutoff: 1×10^{-5} , -outfmt 6), and the OrthoMCL (v.2.0.9) program (percent match cutoff = 30, e-value exponent cutoff = $1e^{-5}$, expansion cutoff = 1.5) (Kramer and Irish, 1999) was then applied to group the BLASTP results into orthologous and paralogous clusters.

The phylogenetic tree for the *L. glauca* genome and other chosen plant genomes was constructed using mutual single-copy orthologous genes. First, MUSCLE (v.3.8.31) software (Edgar, 2004) was used to align the predicted proteins of single-copy genes, and the TRIMAL (v.1.4.rev22) program (<http://trimal.cgenomics.org/introduction>) was used to filter the comparison results (-gt 0.2). Subsequently, a maximum-likelihood phylogenetic tree of single-copy genes was obtained by combining the filtered results using RAxML (v.8.2.10) (model: PROTGAMMAWAG) software (Stamatakis, 2014).

The expansion and contraction of gene families were examined using CAFE (v4.2) software (De Bie et al., 2006) by clustering all homologs in the 15 species, as inferred from the OrthoMCL analysis; this software employed a random birth and death model to estimate the size of each gene family.

Whole-genome duplication and colinearity analyses

The MCScanX package (Wang et al., 2012) with parameters (-a -e 1e-5 -s 5) was used to analyze the syntenic blocks of the genome of four plants after the BLAST (v.2.6.0) alignment (*E-value* cutoff: 1×10^{-5} , -outfmt 6) of protein-coding sequences from these species. For each duplication node in the resulting phylogenetic trees, the program *yn00* in the PAML (v.4.9) (Yang, 2007) package was used to calculate nonsynonymous substitution rates (*Ka*), synonymous substitution rates (*Ks*), and their ratio (*Ka/Ks*) of colinear orthologous gene pairs. Thereafter, the program *mcmctree* (nsample = 1,000,000, burnin = 200000, seqtype = 0, model = 4) in the same package was used to estimate the divergence time of the 15 species based on the phylogenetic tree of these species. In addition, MCScanX was used to identify synteny or colinear segments between the genomes of *L. glauca* and *V. vinifera* and *L. cubeba* and *V. vinifera*, respectively. These results are presented using the ggplot2 (v.2.2.1) program.

TPS gene identification

The proteome dataset of six species were used: *L. cubeba* (Chen et al., 2020c), *Cinnamomum micranthum* (Chaw et al., 2019), *Phoebe bournei* (Chen et al., 2020a), *Persea americana* (Rendón-Anaya et al., 2019), *L. glauca*, and *V. vinifera* (Martin et al., 2010). For *L. glauca*, predicted protein-coding genes were aligned against annotated terpenoid biosynthesis pathway genes in the KEGG database using BLASTP (v.2.6.0) (*E-value* cutoff: 1×10^{-5} , identity > 50%, alignment coverage $\geq 50\%$). For the other five species, all orthologous TPS gene sequences were retrieved from the results of comparative genomics analyses. Branching nodes with bootstrap values of <80% were collapsed.

Metabolomic and transcriptome analysis

To investigate critical genes involved in terpenoid production in *L. glauca*, the content and composition of essential oil in its fruits and leaves at different development stages (fruits: 50, 80, 110, 140, 170, and 200 DAF; leaves: 0, 20, 50, 80, and 140 DAF) were analyzed using GC-MS as described by Sun et al. (2011). Briefly, samples were ground and incubated at 40°C for 30 min. The volatiles were further extracted using SPME fibers with 50/30 μm divinylbenzene/carboxen/polydimethylsiloxane (DVB/CAR/PDMS) (Supelco, Bellefonte, PA, USA). GC-MS analysis was conducted on an Agilent 6890N gas chromatograph coupled to a mass spectrometer (Agilent 5975B, Santa Clara, CA, USA) with a fused silica capillary column (DB-5MS) coated with polydimethylsiloxane (19091 S-433) (internal diameter of 60 m \times 0.25 mm, film thickness of 0.25 μm). We selected the samples of fruits (50, 80, and 140 DAF) and leaves (0, 50, and 140 DAF) at three stages (as mentioned parenthetically) for transcriptome sequencing using the Illumina NovaSeq 6000 platform. Transcriptomic data processing was the same as above 'Plant materials and genome sequencing' section.

QUANTIFICATION AND STATISTICAL ANALYSIS

The quantitative analysis of metabolites used multiple reaction monitoring (Chen et al., 2013), which was determined by total ion chromatogram (m/z). The chromatographic peak area is proportional to the content of the corresponding component. For RNA-Seq data, TPM (Transcripts Per Kilobase of exon model per Million mapped reads) (<https://github.com/lucynwosu/TPM-Transcripts-Per-Million-Normalization-Python>) was used to standardize the count of reads for differential expression of unigenes.

ADDITIONAL RESOURCES

This study does not include additional resources.